Supporting Information for: Statistical variable selection: An alternative prioritization strategy during the non-target analysis of LC-HR-MS data

Saer Samanipour,*,† Malcolm J. Reid,† and Kevin V. Thomas†,‡

†*Norwegian Institute for Water Research (NIVA), 0349 Oslo, Norway*

‡*Queensland Alliance for Environmental Health Science (QAEHS), University of Queensland, 39 Kessels Road, Coopers Plains QLD 4108, Australia*

E-mail: saer.samanipour@niva.no

S1 Experimental Procedures

S1.1 Chemicals

The ACS grade ammonium formate, acetonitrile, and formic acid as well as zinc sulfate were purchased from Sigma-Aldrich, Norway. The QuEChERS extraction powder was obtained from Waters, Norway. The list of all the alkanes artificially added to the background signal is provided in Table S1.

S1.2 Sample Preparation

The sludge samples were extracted employing a QuEChERS method.[1] In short, 1 mL of a 0.1 M zinc sulfate solution was added to 0.5 g of a freeze-dried sludge sample, in order to lysis

Table S1: The number, name, CAS number, and the monoisotopic mass of the alkanes artificially added to the environmental background signal during the semi-synthetic data generation (see section S3).

Number	Compound	CAS	Monoisotopic mass
1	Decane	124-18-5	142.1722
2	Undecane	1120-21-4	156.1878
3	Dodecane	112-40-3	170.2034
4	Tridecane	629-50-5	184.2191
5	Tetradecane	629-59-4	198.2348
6	Pentadecane	629-62-9	212.2504
7	Hexadecane	544-76-3	226.2661
8	Heptadecane	629-78-7	240.2817
9	Octadecane	629-94-7	254.2973
10	Nonadecane	629-92-5	268.3130
11	Pristane	1921-70-6	268.3130
12	Eicosane	112-95-8	282.3286
13	Phytane	638-36-8	282.3286
14	Docosane	629-97-0	310.3600
15	Tricosane	638-67-5	324.3756
16	Tetracosane	646-31-1	338.3913
17	Pentacosane	629-99-2	352.4069
18	Heneicosane	629-94-7	296.3443
19	Hexacosane	630-01-3	366.4225
20	Heptacosane	593-49-7	380.4382
21	Octacosane	630-02-4	394.4539
22	Nonacosane	630-03-5	408.4695
23	Triacontane	638-68-6	422.4851
24	Hentriacontane	630-04-6	436.5008
25	Dotriacontane	544-85-4	450.5164
26	Tritriacontane	630-05-7	464.5321
27	Tetratriacontane	14167-59-0	478.5478
28	Pentatriacontane	630-07-9	492.5634
29	Hexatriacontane	630-06-8	506.5790
30	Heptatriacontane	7194-84-5	520.5947
31	Octatriacontane	7194-85-6	534.6104

cells. To this solution 4 mL of acetonitrile was added for further protein precipitation. 1 g of QuEChERS extraction powder (Waters Milford, MA, USA), including 80-85% Magnesium Sulfate and 15-20% Sodium Acetate, was added to the solution and centrifuged at 2500 rpm for 6 min. The Supernatent was, finally, diluted with 0.9 mL of water and further centrifuged at 3500 G for 10 min. This final extract then was stored in freezer at -20 °C before analysis. The blanks were the extracts of the glassware using the procedure explained above.

S2 Data Treatment

S2.1 Data Binning

All the chromatograms were binned within a width of 10 mDa. This process was performed by generating a vector containing all the measured m/z values in all the samples. During this stage the signal smaller than a preset threshold, in this case 300 counts, is set to zero. The set threshold is defined by the user and based on the data set. For this data set 300 appeared to be a reasonable value once we inspected the level of noise in the chromatograms.

S2.2 Retention Alignment

Retention time alignment is an important step for multivariate statistical tests[2] and a miss-alignment of the variables may have substantial negative effects on the quality of the results.[3] We performed the retention alignment employing a home made algorithm inspired by piece-wise algorithm.[3,4] For the retention alignment, first the user selects a target chromatogram. The target chromatogram should contain several features that are present in all the other chromatograms. A user defined number of locations in the target chromatogram are selected in the next step. The mass spectra of these locations are recorded and are used in order to find the shift needed in the other chromatograms. An inspection window is selected in order to perform the mass spectra correlation. The mass spectra of the target points are correlated to the mass spectra of the same location ± the inspection window. The location

in that window (i.e. the target point ± the inspection window) with correlation coefficient >0.9 and ρ value < 0.05 defines the shift necessary in each chromatogram. In cases that there are multiple locations in the analyzed window which fulfill these criteria, the priority was given to the location with first highest correlation coefficient and the lowest ρ value. This algorithm showed to be effective for alignment of the chromatograms of complex samples.[4–7] More details about this algorithm are given elsewhere.[4] For the alignment of our chromatograms, we employed 30 target locations and an inspection window of 5 scans. In other words, the algorithm could shift the chromatogram by maximum 5 scans (i.e. 2.5 S). For our data set, the largest observed shift was 2 scans in 3 out of 30 target locations of two chromatograms, Figure S1.

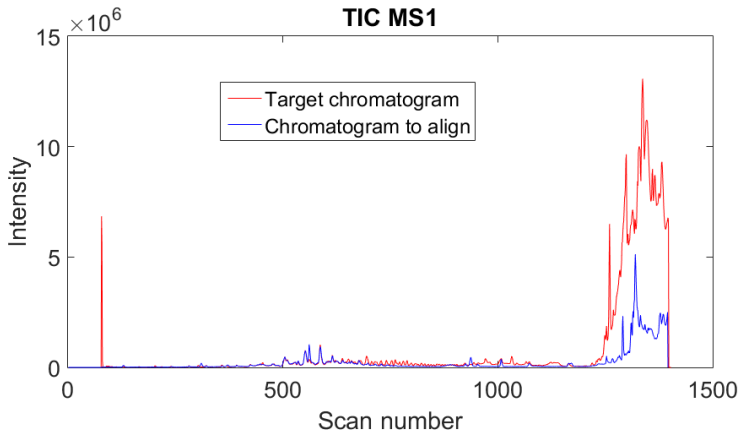

Figure S1: Figure depicting the total ion current (TIC) of the target chromatogram, in red, and the aligned chromatogram in blue.

S2.3 F-ratio calculations

F-ratio or Fisher-ratio is a result of the analysis of the variance in a normally distributed population.[8] The F-ratio is a measure of the observed variability within a group compared to the variability between different groups. Large values of F-ratios indicate that there is more variability between different groups than within each group.[9] As mentioned before, one of the inherent assumptions in the F-ratio calculation is the normal distribution of the data[10]

otherwise a non-parametric test is necessary. A recent study demonstrated that the normal distribution condition can be assumed fulfilled in case of LC-MS data.[11]

Assuming the normal distribution in our data set, we calculated the F-ratio for each independent variable. In order to perform these calculations, a second text file where the sample groups (i.e. row) were specified was submitted to the algorithm. These known classifications enabled us to define the variance within and between groups. Therefore, an F-ratio was calculated for every single variable in the matrix data, Figure S2.

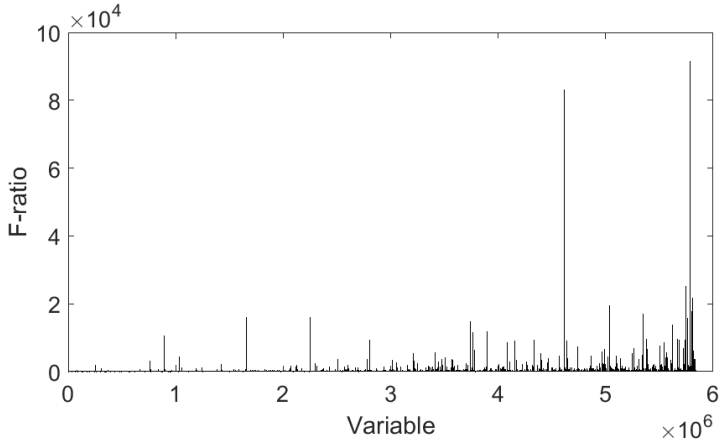

Figure S2: Figure depicting the calculated F-ratio spectra for 6×10^6 variables, which is created using 100 scans of the chromatogram of the sludge sample from Oslo WWTP.

S2.4 Null Distribution Validation and Zero Mask Application

A typical problem when evaluating a large number of statistical hypothesis simultaneously is controlling the rate of false positive discovery.[12] This control is performed via validation. There are different approaches, such as resampling, jackknife, bootstrapping, and permutation.[13,14] Permutation showed to be a reasonable strategy to minimize the rates of false positive discovery when performing evaluation of a large number of statistical hypothesis.[13,15]

In order to perform null distribution validation the F-ratio of each individual variable was

calculated where the components of each group belonged to different group. This was done by rearranging the order of the samples (i.e. rows in the data matrix). This was performed by computing the F-ratio for all possible unique null combinations of the sample arrangements. For example in a case where there are four replicates A1-A4 and four groups A-D, the first unique null combination consists of group one samples A1, B1, C1, and D1 whereas group two comprised of samples A2, B2, C2, and D2 and so on. This approach enables testing the null combination several times and therefore results in a probability distribution of the F-ratios for all the possible unique null combinations, Figure S3. Using the generated null distribution, a probability of false positive discovery may be attributed to each F-ratio. For example in the case of the sludge samples an F-ratio of 28.2 has a probability of 0.05% for false positive detection. This method showed to be effective in simultaneously reducing the rate of false discovery and validation of the F-ratio test.

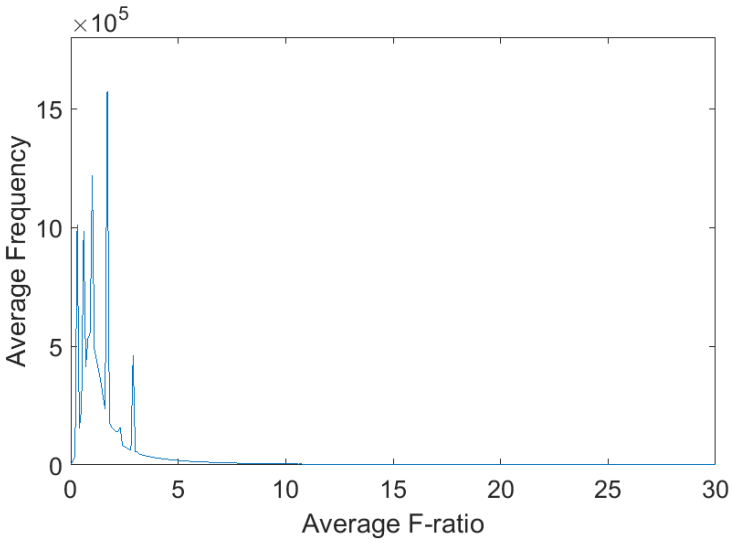

Figure S3: The null distribution calculated for total of 18 samples consisting of 15 sludge samples and 3 blanks.

S2.5 Apex Detection

Even after the zero mask application there may still be some level of redundancy in the data. This implies that there are multiple non-zero variables that can be represented using a unique variable. For example, all the points in a chromatographic peak can be represented by the apex of that peak, which reduces the level of redundancy in the chromatogram. We reduced the level of redundancy in our data through a process here referred to as apex detection. Apex detection consisted of detection of maximums in the both time and m/z domains and group the non-zero points which belong to the same independent variable. During the apex detection we performed the following steps: 1) detection of a maximum point in the extracted ion chromatogram (XIC), 2) finding the baseline in both sides of the maximum point in the XIC, 3) chromatographic peak shape evaluation, and finally 4) the signal to noise ratio (S/N) evaluation.

The maximum point detection is performed in the zero-mask applied chromatogram. Once the maximum point is located, the closest baseline points in both sides of the maximum point are positioned. In order to evaluate the chromatographic peak shape, each point between the maximum point and the baseline are checked to have lower intensity than the neighboring points. In other words, assuming point x_i is the maximum point and x_{i+3} is the baseline point in one side of the maximum point. For x_i to be considered an apex both x_{i+1} and x_{i+2} must have intensities smaller than x_i as well as $x_{i+1} > x_{i+2}$. If a maximum point meets these criteria for both sides, then it can be considered for S/N evaluation. The S/N evaluation enables the removal of the noise from the final feature list. The S/N is evaluated over a user defined window, where the signal (S) is defined as the intensity of the maximum point and the noise (N) is the median of the signal in the binned chromatogram over the selected window. If a maximum point resulted in a S/N larger than the user defined value then that maximum point is considered as a unique feature. This unique feature is considered representative of all the grouped points. For example, all the points related to the maximum

point x_i (i.e. from x_{i-3} to x_{i+3}) are represented with the maximum point. It should be noted that baseline finding, peak shape evaluation and the S/N evaluation are performed on the binned chromatograms while the maximum detection is performed on the zero mask applied chromatogram. There are three main parameters in the apex detection algorithm including max peak width, S/N window, and S/N threshold. The max peak width is a user defined parameter, which sets the distance between the maximum point and the baseline on each side of the maximum point. For our analysis, we used a max peak width of 10 scans, S/N window of 20 scans, and a S/N threshold of 5. These parameters showed to produced the best results during the apex detection, Figure S4. Once the apex detection is performed, a list of m/z and retention time pairs is generated. This list may still have some redundant pairs caused by adducts and isotopes of a certain parent ion. Therefore, adduct and isotope removal should be performed in order to create the final unique feature list.

During our analysis we used a max peak width in the time domain of 4 s, a noise window of 10 scans also in the time domain, and finally a S/N threshold of 5. These parameters were optimized based on the observed performance of the apex detection algorithm when processing the real environmental chromatograms.

S2.6 Adduct and isotope removal

We removed the potential adducts and isotope ions from the m/z and retention time pairs generated from the apex detection. The removal process is performed by looking at a certain mass window after the main ion in the list in order to detect potential m/z values which may be an adduct or isotope of the main ion. Once an m/z value is within the acceptable mass window, then the retention time of that m/z value is compared to the retention time of the main ion. If we observed a match in the expected mass of adduct/isotope and the observed masses as well as the retention time of the main ion and the suspected m/z value, then this m/z value is considered a potential adduct/isotope and is removed from the list in order to

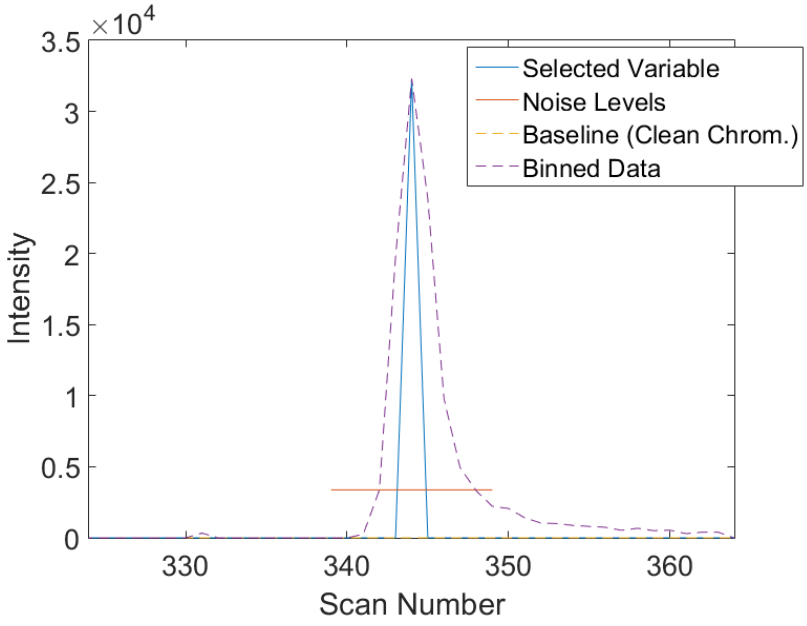

Figure S4: Figure depicting the grouping of three independent non-zero variables, which belong to octadecane, into one unique variable through apex detection process.

create the final unique feature list.

S3 Semi-synthetic Data Generation

We generated a semi-synthetic data set for validation of F-ratio method by combining the background signal coming from the real samples, the synthetic signal of 31 alkanes (Table S1), and added noise. The background signal consisted of the sum of three randomly selected chromatograms out of 15 sludge chromatograms. We generated in total 20 different background signals during each simulation. These background signals were divided in four sample groups with a population of 5 samples for each. The signal of the alkanes was added to these background signals. The alkanes were divided in two groups, consisting in 15 alkanes as true positives and 16 alkanes as true negatives. A true positive was a compound that its between sample group variability was larger than its within sample group variability, which resulted in a large F-ratio for this compound. A true negative, on the other hand, was a chemical where its between sample group variability was smaller than the within sample group variability,

therefore a small F-ratio, Figure S5. Both true positives and true negatives were added into the background signal in the m/z domain having a mass peak width of 30 mDa. Considering the binning width of 10 mDa, each added peak to the background represented three variables with the central variable having the highest intensity. The mentioned mass peak width was selected based on the average observed mass peaks in the environmental chromatograms. Moreover, this enabled us to evaluate the performance of the apex detection algorithm at the same time. The absolute intensity of the central point for the true positives was defined as 5% of the total ion chromatogram (TIC) signal at the addition location, Figure S6. For the locations where the TIC level was smaller than 500, we selected an absolute intensity of the central point of 600. Moreover, some noise was added to the profile of each added alkane in the generated data in order to increase the level of complexity. These low, almost close to the baseline, concentrations enabled us to make sure that the added signal is at an environmentally relevant concentrations, Figure S6. As mentioned before, the alkanes were divided in two groups true positives and true negatives. The true negatives were added at a similar absolute intensity of 1000 $\pm$ 10% for the central point to all the samples, Figure S5. For the true positives, we randomly selected a concentration factor varying from 2 to 8. For example, an alkane may have an average absolute intensity of 1000 $\pm$ 10% in the sample group one, 5$\times$1000 $\pm$ 10% in group sample two, and 4$\times$1000 $\pm$ 10% in group sample three and so on, Figure S5. Both true positives and true negatives were added at randomly selected retention times. They were allowed to partially overlap however their complete overlap was not permitted. We repeated this process for each simulation, which implied the generation of a completely new data set. These data sets then went through all the steps of the F-ratio algorithm in order to detect the added true positives in that data set. It is worth noting that at the end of each simulation a different list of unique features was created.

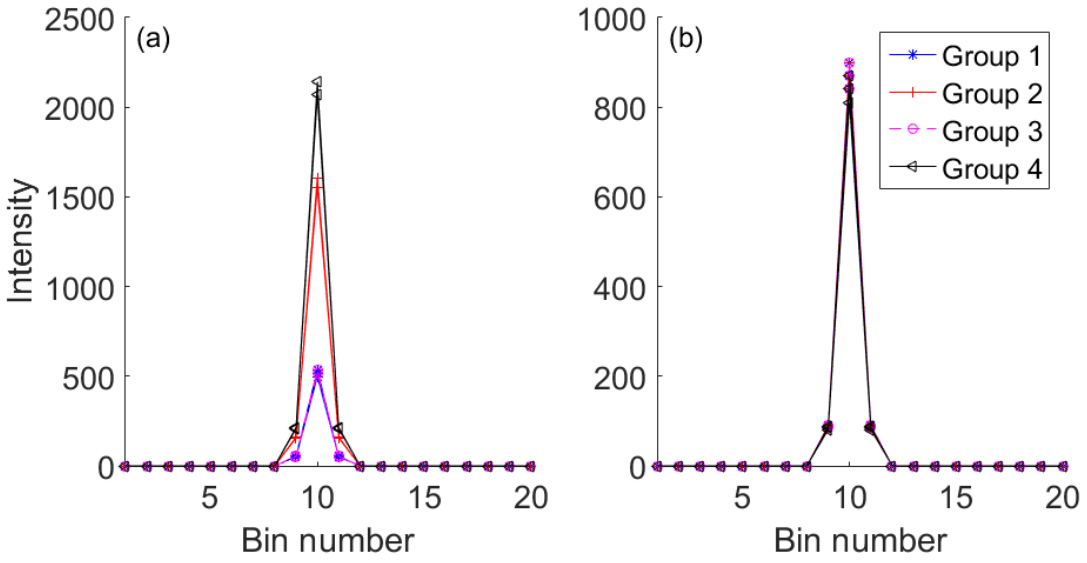

Figure S5: Figure depicting (a) an example of the signal for a true positive added to the background signal and noise, and (b) an example for a true negative added to background signal plus noise.

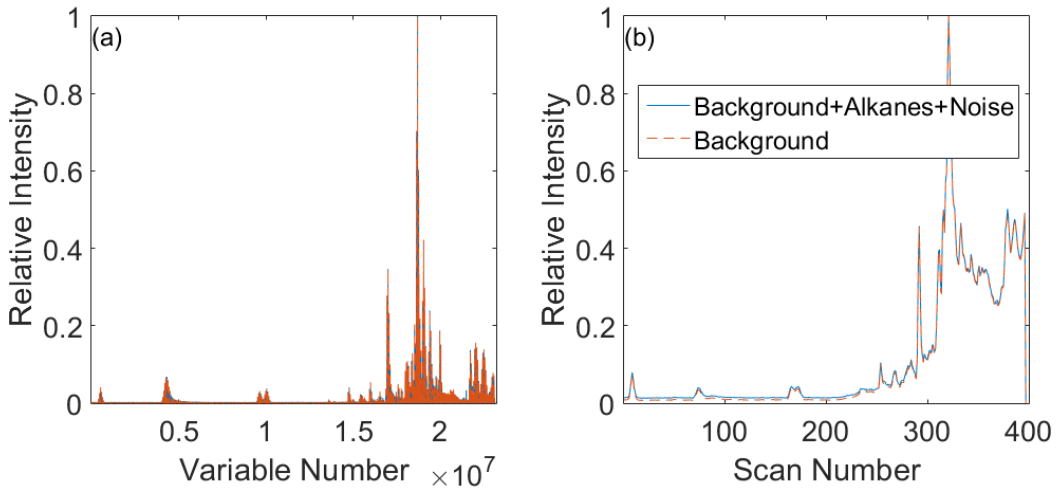

Figure S6: Figure depicting an example of 400 scans of a sludge chromatogram used for (a) showing all the variables before and after adding the alkanes and noise to the background, and (b) the TIC before and after adding the alkanes and noise to the background.

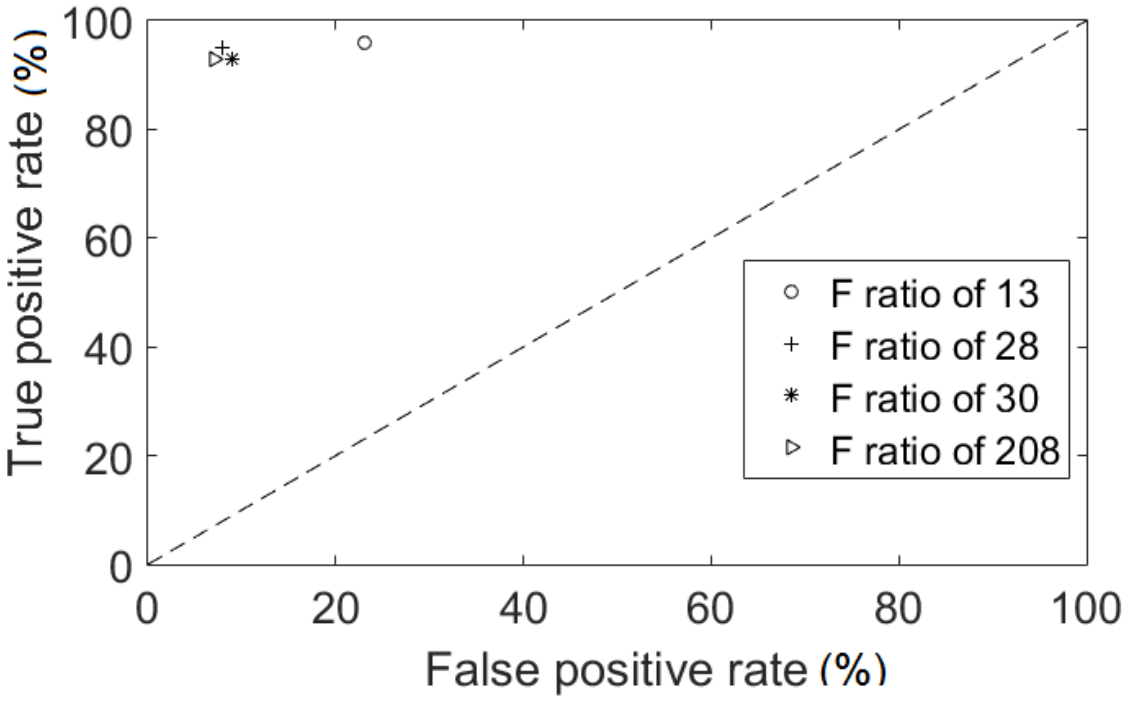

Figure S7: Receiver operating characteristic plot (ROC) for four different F ratio vlaues evaluated.

References

(1) Schenck, F.; Hobbs, J. *Bull. Environ. Contam. Toxicol.* **2004**, *73*, 24–30.

(2) Brereton, R. G. *Applied chemometrics for scientists*; John Wiley & Sons, 2007.

(3) Pierce, K. M.; Hope, J. L.; Johnson, K. J.; Wright, B. W.; Synovec, R. E. *J. Chromatogr. A* **2005**, *1096*, 101–110.

(4) Nadeau, J. S.; Wright, B. W.; Synovec, R. E. *Talanta* **2010**, *81*, 120–128.

(5) Parsons, B. A.; Pinkerton, D. K.; Wright, B. W.; Synovec, R. E. *J. Chromatogr. A* **2016**, *1440*, 179–190.

(6) Pierce, K. M.; Hoggard, J. C.; Hope, J. L.; Rainey, P. M.; Hoofnagle, A. N.; Jack, R. M.; Wright, B. W.; Synovec, R. E. *Anal. Chem.* **2006**, *78*, 5068–5075.

(7) Watson, N. E.; VanWingerden, M. M.; Pierce, K. M.; Wright, B. W.; Synovec, R. E. *J. Chromatogr. A* **2006**, *1129*, 111–118.

(8) Johnson, R. A.; Wichern, D. W.; Others, *Applied multivariate statistical analysis*; Prentice hall Upper Saddle River, NJ, 2002; Vol. 5.

(9) Walpole, R. E.; Myers, R. H.; Myers, S. L.; Ye, K. *Probability and statistics for engineers and scientists*; Macmillan New York, 1993; Vol. 5.

(10) Markowski, C. A.; Markowski, E. P. *Am. Stat.* **1990**, *44*, 322–326.

(11) van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. *BMC genomics* **2006**, *7*, 142.

(12) Benjamini, Y.; Hochberg, Y. *J. R. Stat. Soc.: Ser. B (Methodol.)* **1995**, 289–300.

(13) Vis, D. J.; Westerhuis, J. A.; Smilde, A. K.; van der Greef, J. *BMC bioinf.* **2007**, *8*, 322.

(14) Efron, B.; Tibshirani, R. J. *An introduction to the bootstrap*; CRC press, 1994.

(15) Parsons, B. A.; Marney, L. C.; Siegler, W. C.; Hoggard, J. C.; Wright, B. W.; Synovec, R. E. *Analytical chemistry* **2015**, *87*, 3812–3819.